

PR #33750 完整报告

sgl-project/sglang

[Fix] Inkling works with gs:// runai_streamer paths

合并时间: 2026-08-06 09:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33750>

执行摘要

- 一句话: 修复 Inkling 模型从 gs:// RunAI 路径加载量化配置失败
- 推荐动作: 值得快速阅读, 作为云存储路径接入的修复样板。核心思路是在路径解析时先判断 URI 类型, 再选择下载策略, 代码短小清晰。如果想要深入学习, 可结合 `sglang.srt.utils.runai_utils` 中的 `ObjectStorageModel` 实现, 了解对象存储下载的缓存与生命周期管理。建议关注后续是否补充测试。

功能与动机

PR body 明确说明: Fix Inkling server launch when the model path is a `gs://` URL from cloud storage, loaded with RunAI streamer。即修复 RunAI streamer 场景下模型路径为 `gs://` 时 Inkling 服务无法启动的问题。

实现拆解

实现拆解

1. 引入工具依赖: 在 `python/sglang/srt/models/inkling_common/quantization/config.py` 顶部新增导入 `from sglang.srt.utils.runai_utils import ObjectStorageModel, is_runai_obj_uri`, 复用已有的 RunAI 对象存储辅助函数。
2. 扩展路径解析分支: 在 `_get_raw_quant_config` 函数中, 原逻辑只区分 `os.path.isdir()` 本地目录与 `snapshot_download()` 远程 HF repo 两种情况。新增 `elif is_runai_obj_uri(model_name_or_path)` 分支, 当路径是对象存储 URI 时调用 `ObjectStorageModel.download_and_get_path(model_name_or_path)` 将远程文件下载到本地缓存目录, 再继续后续的 `hf_quant_config.json` 读取。
3. 更新注释与语义: 将原注释调整为“A local path holds hf_quant_config.json directly; anything else resolves its JSON configs first. An object-storage URL is neither a directory nor a valid repo id”, 明确说明对象存储 URL 需要独立处理, 避免误走 `snapshot_download`。
4. 配套测试: 本 PR 未新增或修改测试, 仅依赖 CI (标签 `run-ci`) 验证回归。改动集中在量化配置读取路径, 属于模型加载的公共逻辑, 影响面受控。

关键文件:

- `python/sglang/srt/models/inkling_common/quantization/config.py` (模块 量化配置; 类别 source; 类型 data-contract; 符号 `_get_raw_quant_config`): 量化配置读取函数

`_get_raw_quant_config` 增加对象存储 URI 分支，是本次修复的核心。该函数是 Inkling 模型加载量化配置的入口，修复后 `gs://` 路径才能正常解析。

关键符号： `_get_raw_quant_config`

关键源码片段

`python/sglang/srt/models/inkling_common/quantization/config.py`

量化配置读取函数 `_get_raw_quant_config` 增加对象存储 URI 分支，是本次修复的核心。该函数是 Inkling 模型加载量化配置的入口，修复后 `gs://` 路径才能正常解析。

```
def _get_raw_quant_config(
    model_config: ModelConfig,
) -> dict[str, Any] | None:
    """精简版 weight_utils.get_quant_config，只返回量化配置。"""
    # 优先从 HF config 中读取已有的 quantization_config
    hf_quant_config = getattr(model_config.hf_config, "quantization_config", None)
    hf_text_config = getattr(model_config.hf_config, "text_config", None)
    if hf_quant_config is None and hf_text_config is not None:
        hf_quant_config = getattr(hf_text_config, "quantization_config", None)
    if hf_quant_config is None:
        # compressed-tensors 使用 compressions_config
        hf_quant_config = getattr(model_config.hf_config, "compression_config", None)
    if hf_quant_config is not None:
        return hf_quant_config

    model_name_or_path = model_config.model_path

    # 本地目录直接读；对象存储 URL 需先下载；其余按 HF repo id 快照拉取
    if os.path.isdir(model_name_or_path):
        hf_folder = model_name_or_path
    elif is_runai_obj_uri(model_name_or_path):
        # RunAI streamer 的 gs:// 路径既不是目录也不是合法 repo id
        hf_folder = ObjectStorageModel.download_and_get_path(model_name_or_path)
    else:
        hf_folder = snapshot_download(
            model_name_or_path,
            allow_patterns="*.json",
            local_files_only=huggingface_hub.constants.HF_HUB_OFFLINE,
        )

    # 统一从本地目录读取 hf_quant_config.json
    quant_config_file = os.path.join(hf_folder, "hf_quant_config.json")
    if not os.path.exists(quant_config_file):
        return None
    with open(quant_config_file) as f:
        config = json.load(f)
    return config
```

评论区精华

该 PR 没有实质性的 review 讨论：review_comments_count 为 0，仅有一条作者触发 CI 的评论（/tag-and-rerun-ci），并由维护者 ch-wan 直接批准合并。说明改动简单明确，无需额外质疑。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，主要关注点：
 - 依赖有效性：新增导入依赖 sglang.srt.utils.runai_utils 中的 ObjectStorageModel 与 is_runai_obj_uri，若该模块在部分安装环境中缺失，会导致 ImportError。需要在所有支持 Inkling 的部署环境确认该模块存在。
 - 下载行为：download_and_get_path 可能会触发网络下载，若对象存储凭据或网络不通，可能导致启动失败。但这是 RunAI 场景的固有依赖，且该函数在仓库其他位置已被使用，属于既有能力。
 - 缺少测试覆盖：未新增针对 gs:// 路径的单元测试，后续回归风险略高，建议后续补充。
 - 影响：影响范围较窄：仅影响使用 gs:// 对象存储 URL 作为模型路径的 Inkling 模型部署场景；本地路径和普通 Hugging Face repo id 的加载逻辑完全不变。对现有用户无破坏性影响，但对使用 RunAI streamer 的团队来说修复了启动阻塞问题。代码改动集中在量化配置的公共读取函数，理论上对所有通过该函数加载量化配置的模型都有潜在影响，但分支条件严格限定于对象存储 URI，因此实际影响面可控。
- 风险标记：缺少测试覆盖，新增外部依赖分支

关联脉络

- PR #33108 feat(dgx-spark): add inkling-small MoE support for sm_121: 同属 Inkling 模型支持系列，该 PR 为 Inkling MoE 内核适配 sm_121，本次修复补全了 Inkling 在云存储路径下的加载能力。
- PR #33752 [test] Re-enable a pruned Inkling LoRA unit-test set (68 -> 9 cases): 同为 Inkling 相关改动，该 PR 恢复并裁剪 Inkling LoRA 单测，说明 Inkling 支持正在逐步完善，本修复与其共同组成 Inkling 功能闭环。